Utilizing LLMs for Early-Stage Qualities Elicitation: A Study on Responsible Fact-Checking in Journalism

Roxana L. Quintanilla Portugal[1,5][0000-0001-7693-5353], Lyrene Silva[20000-0003-1772-6062], Henrique P. S. Sousa[3][0000-0003-2150-8113], Eduardo Almentero[4][0000-0002-5664-1080]

[1] Universidad Nacional de San Antonio Abad del Cusco, Perú
[2] Universidade Federal do Rio Grande do Norte, Brazil
[3] Unirio, Brazil
[4] Universidade Federal Rural do Rio de Janeiro, Brazil
[5] Ludwig-Maximilians-Universität (LMU), Germany
roxana.portugal@ifkw.lmu.de, lyrene.silva@ufrn.br,
hsousa@uniriotec.br, almentero@ufrrj.br

Abstract. Requirements elicitation is the process through which engineers interact with information sources to acquire knowledge about a specific domain. In the early stages of software projects, it is uncommon for engineers to elicit non-functional requirements (NFRs) as first-class requirements, as this often demands time and articulation that stakeholders may not readily provide. Large Language Models (LLMs), such as ChatGPT, trained on massive textual datasets, offer a promising opportunity to support this process by generating coherent and context-relevant information about qualities that are key to a given domain problem. In this study, we explore the potential of ChatGPT as an information source for eliciting NFRs related to *responsible fact-checking in journalism*. Using a previously constructed reference model—developed by the first co-author during a series of Design Thinking sessions with journalism professionals—as a gold standard, three requirements engineering experts, none of whom were familiar with the domain, conducted individual chat sessions with ChatGPT and independently constructed Softgoal Interdependency Graphs (SIGs). Our findings go beyond a simple comparison with the gold standard. While some softgoals consistently emerged across sessions (e.g., trust, accuracy, transparency), participants also uncovered quality concerns such as integrity, dignity, and fairness—elements not explicitly included in the original model. These highlight risks that fact-checking practices must proactively mitigate and offer a broader understanding of relevant qualities in the domain. Additionally, the absence of certain softgoals from the LLM-generated models underscores the importance of human–AI collaboration to improve the completeness and contextual richness of SIGs.

Keywords: Non-Functional Requirements, Large Language Models (LLMs), Responsible Fact-Checking, Elicitation.

1 Introduction

Requirements elicitation, especially the elicitation of NFRs, is crucial in the field of requirements engineering. For years, several authors have pointed out that NFRs should

Proceedings of the 28th Workshop on Requirements Engineering (WER2025),
August 20-22, 2025, Rio de Janeiro-RJ, Brazil.
https://doi.org/10.29327/1588952.28-29

be considered as first-class requirements. However, there are few methods to elicit them. One of the most explored methods is the NFR framework proposed by Chug et al. [1], which in its notation proposes to model NFRs with a top-down strategy, meaning starting from the qualities of desired software to decompose the abstraction level of a quality until reaching operationalizations that allow its satisfaction.

Although this notation has facilitated the creation of the models called SIG, which are reusable as demonstrated by the transparency catalog [2], they are still scarce. Furthermore, creating catalogs assumes that the requirements engineer has obtained information from stakeholders or other sources. Herein lies the central problem: stakeholders do not always find it easy to articulate qualities, which is why catalogs support them.

Textual information sources are invaluable for constructing catalogs. Historical data from one or multiple projects within a domain, as well as requirement-related content available online [3], such as GitHub projects or application reviews [4][5], offers rich insights into user needs, making it possible to identify NFRs. Although explicitly stated NFRs may be few, Big Data analysis can yield sufficient instances to anticipate the necessary qualities in a specific domain.

While these textual sources have been widely explored using NLP techniques in requirements engineering, primarily for extracting and analyzing functional requirements, working with large volumes of unstructured data typically demands extensive preprocessing and considerable computational resources. In this context, LLMs offer a promising alternative to support the retrieval of relevant information about domain-specific qualities. However, a key challenge in relying on LLMs lies in ensuring the reliability, consistency, and contextual relevance of the information they generate.

In this study, we explore the potential of LLMs, specifically ChatGPT, as a source of domain knowledge to support the early elicitation of NFR related to *responsible fact-checking in journalism*. This practice refers to the ethical and systematic verification of information before its dissemination, and it plays a central role in combating misinformation and preserving the credibility of news organizations. However, due to its domain-specific nature, it can be difficult for requirements engineers, especially those without prior knowledge of journalism, to elicit relevant NFRs in the early stages of a project. In the context of responsible fact-checking, the expected NFRs include qualities such as accountability, inmediacy, newsworthiness, transparency, trust, compliance, care and completeness, which are crucial for ensuring ethical and effective verification practices

To investigate this, three requirements engineering specialists with extensive experience in NFR modeling participated in the study. Notably, none of them had prior knowledge of the journalistic context, which allowed us to examine how effectively an LLM can assist experts in approximating the knowledge embedded in a domain-specific model. This reference model, constructed over several months by the first co-author familiar with the domain, is the gold standard for comparison.

This paper is structured as follows: Section 2 reviews related work. Section 3 details the protocol followed by three requirements engineering experts to elicit information from ChatGPT. Section 4 presents the results by comparing the resulting SIGs with the gold standard. Section 5 discusses the insights gained from this exercise, and section 6 concludes this paper and outlines directions for future research.

2 Related Works

Ronanki et al. [6] discuss the potential, limitations, and challenges of using LLMs to generate requirements. The study compares requirements generated by ChatGPT and humans, and its results demonstrate that ChatGPT produces high-quality requirements (highly rated in terms of being Abstract, Atomic, Consistent, Correct, and Understandable) despite ambiguity and feasibility.

Chen et al. [7] define prompts to generate goal models using GPT-4 and evaluate such prompts through experiments. The results obtained are promising, considering the knowledge retention of GPT-4, although the modeled elements are abstract and have syntactic and semantic errors.

Marczak-Czajka and Cleland-Huang [8] define prompts to generate user stories with human values as creative triggers for stakeholders in the requirements elicitation process. An experimental study with students was conducted, and its findings showed the potential of GPT for this activity. GPT can facilitate eliciting and specifying well-structured and meaningful human value stories for software products.

Although our paper also explores the potential of LLMs, it differs from prior works [6][7][8] in that we use ChatGPT to support experts in eliciting knowledge for softgoal modeling rather than relying on the LLM to generate softgoal models, requirements, or user stories directly. To the best of our knowledge, few studies have investigated LLM-assisted modeling [9][10], and none have specifically addressed the modeling of NFRs using knowledge extracted from LLMs. Another distinction is that we are not (yet) focused on prompt engineering to maximize information retrieval; instead, we analyze whether ChatGPT can provide relevant and meaningful insights regardless of the specific prompt used. This positions our work within a different scope—less about generating complete artifacts and more about uncovering foundational quality attributes to inform the modeling process.

3 Case Description

This study investigates how LLMs, particularly ChatGPT, can assist requirements engineering experts in eliciting NFRs for responsible fact-checking in journalism. The case focuses on comparing expert-generated SIGs with a pre-established gold standard, developed over several months through Design Thinking workshops with journalists.

Participants were asked to manually construct a SIG addressing quality concerns relevant to journalistic fact-checking based on the information gathered from ChatGPT. While no specific elicitation or modeling method was imposed, a set of constraints was defined to ensure consistency across their approaches:

- They were restricted to using only two information sources: the German Press Code [11] which was provided to participants as a static reference document (not uploaded to ChatGPT), and ChatGPT 3.5 accessed through a free account. ChatGPT was selected due to its broad accessibility and the absence of domain-specific fine-tuned alternatives readily available at the time of the study.
- Each participant was allowed to ask ChatGPT a maximum of seven questions. The limit of seven questions per participant was set to simulate the constraints

typically faced in early-stage elicitation when time or access to domain experts is limited.
- The resulting SIGs were not required to include operationalizations.

It is important to note that we chose ChatGPT in its base form because its web-scale training captures both formal guidelines and real-world discussions about fact-checking qualities. This broad coverage contrasts with specialized corpora, which focus only on normative aspects, and allowed us to explore whether an out-of-the-box LLM could surface relevant qualities without fine-tuning or RAG (Retrieval-Augmented Generation).

Once the SIGs were completed, each participant filled out a comparison table to identify similarities and differences with the reference model. This allowed us to analyze how well a language model can help experts approximate the knowledge embedded in a domain-specific model developed over several months by the fourth co-author. This gold standard was created through Design Thinking sessions in a newsroom setting [12], where journalists, editors, and social media managers articulated their need for a tool — potentially AI-assisted — to conduct responsible fact-checking.

4 Results

The artifacts generated in this experiment include three chat transcripts with ChatGPT and three SIGs on Responsible Fact-Checking [13].

As this paper focuses on exploring the potential of LLMs to support the early elicitation of non-functional requirements, the author of the gold standard reviewed each SIG produced by the other participants. This review aimed to identify which softgoals were absent from the gold standard and to reflect on why certain qualities may have been overlooked or emphasized differently across the various SIGs.

4.1 NFR Models

Table 2 compares the three SIGs generated by the participants and the reference model. Each model reflects a different elicitation strategy and interpretation of the information retrieved through ChatGPT. Table 3 provides a side-by-side comparison of the softgoals identified in the participant SIGs and the gold standard, highlighting overlapping and novel elements.

4.2 Key Contributions and Gaps Identified in Participant SIGs

The SIGs generated with ChatGPT successfully captured a significant portion of the domain knowledge, especially regarding Accuracy, Transparency, and Trust. Notably, Transparency was represented in the participants' models through refined subtopics such as Process, Methods, and Sources. These distinctions are valuable as they can facilitate the auditing of the fact-checking process and contribute to strengthening reader trust. The relationship between Trustworthy and Solidibility also emerged as important. The latter provides a stronger foundation for trust by emphasizing that evidence must be traceable, replicable, and well-founded.

Table 2. SIGs Comparison

Gold[1]	P1	P2	P3
├── Feeling of Success [journalist's professional experience] │ └── Immediacy [fact-checking responsiveness] │ ├── Trust [audience perception of fact-checking] │ │ ├── Clear-sighted trust [fact-checkers judgment] │ │ ├── Completeness [fact-checking effort] │ │ │ ├── Confidence [fact-checking process] │ │ │ │ └── Stability [journalistic work] │ │ │ │ └── Valuable [fact-checker's effort] │ │ │ │ └── Successful [impact on public] │ │ │ └── Care [involved parties] │ │ └── Transparency [fact-checking process] │ │ └── Accuracy [information verified] │ ├── Newsworthiness [fact-checking criteria] │ ├── Reliable Sources [journalistic input sources] │ ├── Truthful Sources [journalistic input sources] │ └── Compliance [legal and ethical standards] │ ├── Accountability [Journalistic responsibility] │ ├── Accountability [Personal] │ ├── Accountability [Company] │ └── Completeness [fact-checking effort] │ ├── Confidence [fact-checking process] │ │ └── Stability [journalistic work] │ │ └── Valuable [fact-checker's effort] │ │ └── Successful [impact on public] │ └── Care [involved parties]	Credibility [Press] ├── Responsible [use of info] ├── Responsible [correction] ├── Responsible [fact-checking] │ ├── Fairness │ │ ├── Respectful [diversity] │ │ ├── Respectful [human dignity] │ │ └── Respectful [privacy] │ ├── Objectivity │ │ ├── Unbiased │ │ │ ├── No sensationalism │ │ │ └── No personal or political bias │ │ └── Balanced [multiple views] │ ├── Accuracy [info] │ │ ├── Reliable [info] │ │ ├── Reliable [sources] │ │ └── Trustworthy [info] │ ├── Transparency │ │ ├── Transparency [process] │ │ ├── Transparency [sources] │ │ ├── Transparency [editings] │ │ └── Transparency [sanctions] │ ├── Accessibility │ │ ├── Understandable [language] │ │ └── Availability [format] │ ├── Independence │ │ ├── Independence [editorial] │ │ └── Independence [financial] │ └── Timely │ ├── Efficience [process] │ └── Efficience [cross-referencing]	Responsible [Fact-checking] ├── Accuracy │ ├── Thoroughness │ ├── Impartiality │ ├── Independence │ └── Accountability ├── Timeliness │ ├── Efficiency │ ├── Readiness │ ├── Prioritize │ ├── Availability │ └── Collaborative ├── Auditability │ ├── Evaluability │ ├── Traceability │ ├── Clarity │ ├── Accessibility │ └── Consistency │ ├── Ethically │ ├── Uniformity │ └── Objectivity	Responsibility [Fact-checking] ├── Accuracy [Fact-checking] ├── Impartiality [Fact-checking] ├── Solidibility [Evidence] ├── Transparency [fact-checking] │ ├── Transparency [Process] │ ├── Transparency [Methods] │ └── Transparency [Sources] ├── Integrity [Fact-checking] │ ├── Integrity [Involved parties] │ │ ├── Fairly [Fact-checked conclusions] │ │ └── Accuracy [Fact-checked conclusions] ├── Dignity [involved parties] │ ├── Fairly [Fact-checked conclusions] │ └── Accuracy [Fact-checked conclusions]

[1] SIG model available upon request from the corresponding author.

Table 3. How Participant Models Complement and Expand the Gold Standard for Responsible Fact-Checking

Part.	Softgoal in participants	Present in Gold Standard?	Comment on Possible Integration
P1	Credibility [Press]	✓ (as Trust)	Functionally equivalent to Trust but with institutional focus
	Responsible [use/correction/fact-checking]	✓ (partially)	Reflected in Accuracy [information verified]
	Fairness	✗	Contributing to Care
	Respectful [diversity/dignity/privacy]	✗	Subdimensions of Care
	Objectivity, unbiased, no sensationalism	✗	Contributing to Clear-sighted trust
	Balanced [multiple views]	✗	Contributing to Clear-sighted trust
	Trustworthy [info]	✓ (different focus)	Complements Trust [audience perception]. Trust can be decomposed in both.
	Transparency [sources/editings/etc.]	✓ (partially)	Decomposing Transparency [fact-checking process]
	Accessibility, understandable [language], Availability [format]	✗	Contributing to Transparency, as supported by TSIG [2]
	Independence [editorial/financial]	✗	Contributing to Auditability as internal guarantees
	Timely, efficience [process/cross-referencing]	✓ (as Immediacy)	Functionally equivalent; different in expression
P2	Responsible [Fact-checking]	✓	Is the central quality but not explicit in Gold Standard SIG
	Accuracy	✓	
	Thoroughness	✓ (partially)	Could be mapped to Completeness [fact-checking effort]
	Impartiality	✓ (partially)	Could be mapped to Clear-sighted trust or Care
	Independence	✗	Contributing to Auditability as internal guarantees
	Accountability	✓ (partially)	Contributing to Accuracy reinforces the fact-checker's commitment to truthfulness.
	Timeliness	✓ (as Immediacy)	Could be mapped to immediacy
	Efficiency	✓ (as Immediacy)	Could be mapped to immediacy
	Readiness	✓ (as Immediacy)	Contributing to Immediacy
	Prioritize	✓ (as Valuable)	Can be mapped to Valuable [fact-checker's effort]
	Availability, accessibility	✗	Contribute to Transparency, as in TSIG [2]
	Collaborative	✗	Contributing to Clear-sighted trust guarantee Objectivity
	Auditability, evaluability, traceability, clarity	✓ (partially)	Contributing to Transparency as in TSIG [2] can reinforce the fact-checking process
	Consistency, ethically, uniformity, objectivity	✗	Contributing to Confidence
P3	Responsible [Fact-checking]	✓	Is the central quality but not explicit in Gold Standard SIG
	Accuracy [Fact-checking]	✓	
	Impartiality [Fact-checking]	✓ (partially)	Contributing to Clear-sighted Trust or Care
	Solidibility [Evidence]	✓ (as Reliable Sources)	Solidibility, contributing to Reliable Sources, demands that evidence be traceable, replicable and well-founded.
	Transparency [Fact-checking]	✓	
	Transparency [Process/Methods/Sources]	✗	Subdimensions of Transparency
	Integrity[Fact-checking]	✗	Contributing to Care
	Fairly [Fact-checked conclusions]	✗	Contributing to Integrity
	Accuracy [Fact-checked conclusions]	✗	Accuracy may be decomposed in Accuracy [source/conclusion]
	Dignity [involved parties]	✗	Contributing to Care

In addition to capturing core qualities, the participants' SIGs introduced several new softgoals that could enrich the gold standard. Fairness (P1, P3), Objectivity (P1), and Independence (P2) were proposed as fundamental to the ethical foundation of fact-checking, directly influencing public perception and acceptance of verified content.

P1and P2 also emphasized process- and norm-oriented softgoals such as Auditability, Traceability, Clarity, Evaluability, Understandability, and Accessibility, which provide internal guarantees for responsible verification practices. Many of these qualities are already organized and supported in the TSIG framework [2].

P3, in turn, introduced ethical and human-centered values, including Integrity, Dignity, and Fairness, highlighting the importance of respectful treatment of individuals involved in fact-checking processes.

However, some key softgoals from the gold standard were missing or underrepresented in the participants' models, such as Feeling of Success, Completeness, and Valuable [fact-checker's effort]. This highlights the experiential and context-driven nature of the gold standard, which was shaped through co-design sessions with journalists as they articulated their expectations for what a Responsible AI tool should provide in the fact-checking process.

The incorporation of new softgoals identified through LLM-supported elicitation, such as fairness, objectivity and independence, warrants further discussion. While these qualities are relevant, their integration into the gold standard requires careful validation to ensure they align with domain-specific priorities and are not merely general ethical ideals detached from journalistic practice.

5 Conclusion

The results highlight the inherently subjective nature of NFR catalogs and reinforce the need to involve both NFR experts and domain specialists. This study shows that completeness can be improved by comparing and complementing different model versions, as done here with the three participant-generated SIGs.

While the Design Thinking workshops followed a user-centered approach with news professionals, they may have overlooked the expectations of news consumers, such as demands for transparency. This aligns with Michael Jackson's insight that understanding requirements involves engaging with the problem world, not just the system-to-be[14].

Future work in journalism could explore integrating LLMs into co-design sessions, where their suggestions can be refined through human input. Furthermore, LLMs could help not only elicit high-level qualities but also propose operationalizations, supporting the balancing of competing demands like accuracy and timeliness in news production[15].

In requirements engineering, future research could focus on methodological aspects, such as refining prompts and collaborative review processes, to develop more systematic approaches for integrating LLMs into NFR elicitation workflows. This study highlights the potential of LLMs to support early NFR elicitation and the critical role of human judgment in ensuring contextual relevance and completeness.

Acknowledgements

We extend our sincere gratitude to the Volkswagen Foundation for sharing data, and Portugal appreciates the support of the research project "Towards Responsible AI in Local Journalism" with grant number 9B761 and 9B390.

References

1. Chung, L., Nixon, B.A., Yu, E., Mylopoulos, J.: *Non-Functional Requirements in Software Engineering*. Springer, Boston, MA (2000). doi: 10.1007/978-1-4615-5269-7.
2. J. C. S. do P. Leite and C. Cappelli, "Software Transparency," *Bus Inf Syst Eng*, vol. 2, no. 3, pp. 127–139, Jun. 2010, doi: 10.1007/s12599-010-0102-z.

3. Dewi, R., Mutia, –, Raharjana, I.K., Siahaan, D., Fatichah, C.: Software requirement-related information extraction from online news using domain specificity for requirements elicitation. In: Proceedings of the 2021 10th International Conference on Software and Computer Applications, pp. 81–87 (2021).https://doi.org/10.1145/3457784.3457796
4. Portugal, R.L.Q., Casanova, M.A., Li, T., Leite, J.C.S.P.: GH4RE: Repository recommendation on GitHub for requirements elicitation reuse. In: *CAiSE-Forum-DC*, pp. 113–120 (2017). https://ceur-ws.org/Vol-1848/CAiSE2017_Forum_Paper15.pdf
5. Liu, Y., Liu, L., Liu, H., Yin, X.: App store mining for iterative domain analysis: combine app descriptions with user reviews. Softw. Pract. Exp. **49**(6), 1013–1040 (2019). https://doi.org/10.1002/spe.2693
6. Ronanki, K., Berger, C., Horkoff, J.: Investigating ChatGPT's potential to assist in requirements elicitation processes. In: 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2023), pp. 354–361. IEEE, New York (2023). https://doi.org/10.1109/SEAA60479.2023.00061
7. Chen, B., Chen, K., Hassani, S., Yang, Y., Amyot, D., Lessard, L., Mussbacher, G., Sabetzadeh, M., Varró, D.: On the use of GPT-4 for creating goal models: an exploratory study. In: 31st IEEE International Requirements Engineering Conference Workshops (REW 2023), pp. 262–271. IEEE, New York (2023). https://doi.org/10.1109/REW57809.2023.00052
8. Marczak-Czajka, A., Cleland-Huang, J.: Using ChatGPT to generate human-value user stories as inspirational triggers. In: 31st IEEE International Requirements Engineering Conference Workshops (REW 2023), pp. 52–61. IEEE, New York (2023). https://doi.org/10.1109/REW57809.2023.00016
9. Ferrari, A., Abualhaijal, S., Arora, C.: Model generation with LLMs: from requirements to UML sequence diagrams. In: 32nd IEEE International Requirements Engineering Conference Workshops (REW 2024), pp. 291–300. IEEE, New York (2024). https://doi.org/10.1109/REW61692.2024.00044
10. van Nifterik, S.: Exploring the potential of large language models in supporting domain model derivation from requirements elicitation conversations. Master's thesis, Utrecht University (2024).
11. German Press Council: German Press Council. [Online]. Available: https://www.presserat.de/en.html. Accessed: 7 Apr 2025.
12. Portugal, R.L.Q., Wilczek, B., Eder, M., Thurman, N., Haim, M.: Design thinking for journalism in the AI age: towards an innovation process for responsible AI applications. In: Proceedings of the Joint Computation + Journalism and European Data & Computational Journalism Conference 2023, Zurich, Switzerland, 22–24 June 2023. https://openaccess.city.ac.uk/id/eprint/30698/
13. Portugal, R.L.Q.: Utilizing LLMs for early-stage qualities elicitation: a study on responsible fact-checking in journalism. *Zenodo* (2025). https://doi.org/10.5281/zenodo.15192124
14. Jackson, M.: Problem Frames: Analyzing and Structuring Software Development Problems. Addison-Wesley, Harlow (2001)
15. Portugal, R.L.Q., Delle Ville, J., Antonelli, L.: Implementing accuracy for responsible AI in newsrooms. In: Proceedings of the 27th Workshop on Requirements Engineering (WER 2024), Buenos Aires, Argentina (2024). https://doi.org/10.29327/1407529.27-30